Human Spatio-Temporal Attention Modeling Using Head Pose Tracking for Implicit Object of Interest Discrimination in Robot Agents

Corey Johnson and Lynne E. Parker

***Abstract*— Hazardous search missions are an excellent application domain for human-robot teams because cost-effective robot systems could be leveraged to reduce total mission duration time, improve search space thoroughness, and reduce human exposure to danger. Efficiently pairing robotic agents with human workers requires leveraging implicit communication when explicit techniques are either unavailable, socially unnatural, or impractical, such as is often the case with challenging search and rescue missions. Although successful implicit communication methods for robotics systems exist, (e.g., gestural recognition, activity recognition, and gaze management), a full suite of effective natural communication skills remains an open problem in robotics, thus preventing human-robot team solutions from being more commonplace. To help address this capability gap, we introduce a technique to implicitly model human spatio-temporal attention with a 3D heat map based on head pose trajectory tracking. We then show that this version of attention modeling can be applied by a robot agent to reliably extract Object of Interest (OOI) information for use in improving implicit communication in human-robot teams. This technique is evaluated in an OOI search task and a shared workspace clustered OOI discrimination task.**

***Index Terms*— Human Attention Modeling, Human Robot Teams, Head Pose Tracking, Spatio-Temporal Attention Maps.**

I. INTRODUCTION

A current open problem for human-robot teams is over-reliance on explicit communication, which can result in slow, unnatural interactions or communication overload among team members. Additionally, for challenging real-world tasks such as search and rescue, the quality and frequency of successful explicit communication messages can be inhibited by noisy environments, distance between peers, large team sizes, and mission urgency. Clearly, natural interaction for effective teamwork in challenging conditions requires both explicit and implicit communication.

Current techniques for implicit communication in robotics include gesture recognition [1], activity recognition [2], and gaze management [3]; however, a full suite of skills necessary for clear and natural communication in human-robot teams is still an open issue. For example, robots typically used for hazardous duty in bomb squad operations still require significant amounts of explicit teleoperation. This reliance on explicit expert control makes it difficult to increase the number robots on a team, slows down mission operations, and can preclude the use of robot agents for time-critical tasks. This open problem presents the technical challenge of how to apply available mobile sensors and computation resources to improve natural communication skills and situational awareness in robot agents.

Fig. 1. An observing robot uses STAM-Heat to help search for OOIs.

While visual attention modeling has been studied for decades [4], most of this prior work has focused on 2D visual attention [5]. More recent work relevant to robotics has been studying this challenge for 3D applications, although the available techniques are still limited. For example, Potapova, et al. [6], point out a number of open challenges for 3D attention modeling, such as how an integrated attention system should smoothly switch between individual attentional mechanisms for sub-tasks since different techniques each have unique strengths for a variety of usage contexts. Our paper does not attempt to propose a comprehensive 3D attention model that addresses all these challenges. Instead, we focus on one specific question in the context of attention modeling for robotics – namely, how can a robot know what object a human teammate is focused on in situations in which eye gaze is not available, but head pose is available? We believe that this is an important step in implicit human-robot teaming, since knowledge of the human's objects of interest can provide meaningful clues to the robot in how to be a collaborative teammate.

Building on the successes of these previous techniques, we extrapolate the 2D spatio-temporal attention map concepts into the 3D domain in order to facilitate implicit modeling of human subject attention for broadening the natural communication skill set in robotics. We introduce three variations of Spatio-Temporal Attention Models (STAMs)

This work is supported in part by the National Science Foundation under Grant No. IIS-1427004.

Corey Johnson and Lynne Parker are with the Distributed Intelligence Laboratory in the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA, {cjohn221,leparker}@utk.edu

based on head pose trajectory tracking and also introduce a STAM Gradient Slicing (GS) technique for clustered OOI discrimination. Figure 1 shows an example of a STAM rendered in 3D space based on a subject's head pose. STAMs allow an Observing Agent (OA) to passively view a human teammate and temporally model the human's field of view in 3D space, which can facilitate an implicit understanding of the teammate's set of information, items of interest, current goals, and probable future actions.

In the three STAM variations, which were chosen based on common usage in 2D domains for eye tracking applications, STAM-Heat uses thermal mapping, STAM-Tmin use time-minimal immediate mapping and STAM-Basic uses normal spatio-temporal mapping. These techniques are evaluated in an Object of Interest (OOI) search task and a shared workspace clustered OOI discrimination task.

II. RELATED WORK

Implicit communication via human subject attention modeling is common in the fields of psychology, marketing, and computer user interfaces, but is typically relegated to 2D domains such as computer monitor screens. Typical 2D visual modeling techniques are derived from tracking the subject's eye gaze or body pose and thus rely critically on the accuracy of these tracking systems. Hence, we first overview related work in implicit communication techniques and then discuss typical pose tracking technologies used in generating spatio-temporal human attention models.

A. Implicit Communication Techniques

Heat Mapping of Eye Gaze (HMEG) within monitor screen coordinates is a common implicit attention modeling technique that allows a subject's gaze to be mapped over time in order to model areas of attention within the 2D screen space. HMEG techniques have been used to help discover stages of child developmental psychology and have also been applied to determine the effectiveness of webpage information layout for optimizing content design. In the field of marketing analysis, for example, the company Package In Sight [7] has used a gaze tracking system to analyze the effectiveness of product packaging designs for items on store shelves. The success of these HMEG attention modeling techniques for implicit communication inspires the 3D techniques used in our STAM implementations.

The implicit communication technique of perspective modeling in robotics is commonly used for tasks that require cooperative task planning, peer attention, and OOI discrimination. For example, these works include models for peer joint attention [8] and OOI occlusions in peer perspective taking for reference resolution [9]. Passive acquisition of human activity classification [2] and gesture recognition [1] are also key skills in implicit communication for robotics applications. These techniques allow robotic systems to implicitly understand current human activities, be directed via natural intuitive means, and effectively plan for human interaction or collaborative tasks. They have also been successful for improving human-robot interaction but leave challenging gaps open for the problem of implicit communication in robotics. For example, in human-robot team search missions, these current techniques could be used to automatically recognize that a certain type of search activity has begun, then be gesturally directed to gaze in a certain location or possibly take perspective reference for passively understanding a human's set of information. However, they do not fully provide a spatio-temporal model of the subject's attention in 3D space, or a map of the human's searched space, or sufficiently assist with OOI discrimination, or with mapping of these item's location and discovery state.

B. Tracking Technologies for Building Attention Models

Attention modeling techniques typically rely on specialized sensor systems to track eye gaze or body pose. For mobile robot applications, a self-contained tracking system that works from a single perspective while providing a high degree of head pose orientation range such that subjects can be tracked from the aft perspective is desired. These features are not currently found in low-cost commercial systems.

The TobiiPro is popular for passively tracking eye gaze within 2D monitor coordinates, whereas natural human interface sensors such as the PrimeSense can provide 3D human skeleton pose tracking by using a single depth sensor camera. Multi-camera arrays for human motion capture, such as those used for cinema special effects and sports video games, can provide accurate 3D skeletal pose extraction by tracking a specially marked bodysuit. Other techniques include facial feature matching of eyes and nose in order to orient head pose. Additionally, 2D skeletal pose can be extracted from a video camera stream using OpenPose. A survey [10] on head pose estimation techniques in computer vision includes further information on alternative head pose tracking techniques.

Each of these tracking techniques have advantages and disadvantages for metrics such as pose range, accuracy, sensor distance, number of sensors required, processing speed, and specific pose elements provided. For example, most of the feature based matching systems cannot accurately track the rear of a human head, which significantly limits the effective orientation range of the attention modeler. Additionally, the multi-camera motion capture systems are not typically practical for single perspective OAs such as those used in mobile robotics teams. Systems such as OpenPose would require processing two stereo image streams at a high frame rate and are also not specifically focused on providing accurate head pose orientation in 3D space. These unique characteristics of each tracker motivate their selection for use in different attention modeling applications.

Due to the specific limitations of these common tracking systems to work from a single perspective while providing a high degree of head pose orientation range, this work uses two different custom head pose tracking systems for the included experiments. These advantages of these trackers are further discussed in the approach section.

III. APPROACH

In this work, the fundamental approach to facilitating implicit communication in human-robot team applications uses the method of 3D STAMs built from tracking human head pose trajectory. The OA uses this model for extracting probable OOIs and candidate search spaces based on relative attention zones within the environment. We first discuss our approach for building STAMs in 3D space and then discuss the gradient slicing technique of dynamically scoping ROIs for OOI discrimination. This section then discusses the model verification techniques used for validating our STAM implementation with the head pose tracker.

The overall system diagram is shown in Figure 2. The inputs to the system are the RGB-D camera and head pose. The outputs are the human subject's STAM with the OOI types, locations, and labeled discovery states to indicate if the OA thinks the human has observed the OOIs yet.

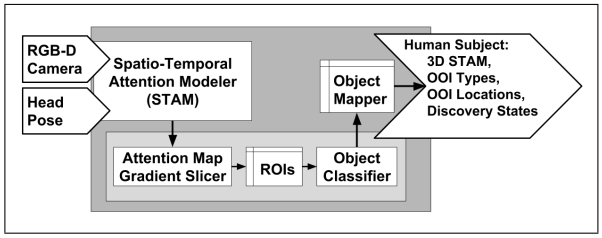

Fig. 2. System Diagram.

A. Spatio-Temporal Attention Mapping in 3D Space

The STAMs in this work are built by modeling the human subject's head pose trajectory in 3D space over time. The OA achieves this by using a PrimeSense depth camera sensor to scan the environment and then project the subject's volumetric area of gaze envelope into it for calculating the model of attention. The gaze envelope shape selected for this work uses a simple cylindrical model; however, other envelope shapes such as dual ellipsoid cones could also be used with this technique. The gaze envelope size and attention score rates are flexible hyper-parameters as discussed in the experiments section.

As shown in Algorithm 1, areas within the subject's gaze envelope accumulate attention value scoring in the model over time and when the subject's gaze envelope changes locations in the environment, the model reduces the attention scoring of older locations no longer found within the current gaze envelope.

Algorithm 1 Update STAM

```
Given a new RGB-D image and head pose:
For all 3D depth pixels p:
  If p exists inside current head pose gaze envelope:
    Determine gaze vector centerline distance
    Increase attention score for p
  Else:
    Decrease attention score for p
Interpolate new attention scores in the output model
```

The three STAM variations in this work use different temporal and attention score interpolation schemes. STAM-Tmin simply uses the immediate gaze envelope and does not accumulate attention over time. STAM-Heat interpolates the attention scores over time and tapers off the attention scoring near the periphery of the gaze envelope. STAM-Basic builds a simple spatial attention map over time and does not use heat map interpolation.

B. Object of Interest Discrimination

The constructed model of the subject's historical areas of attention allows the OA to extract the probable OOIs based on relative attention scoring within the environment. OOI discrimination is achieved by passing a parsed Region of Interest (ROI) from a STAM to an Object Classifier (OC). The Gradient Slicing (GS) technique, as shown in Algorithm 2, can be paired with STAM-Heat by stripping out increasingly larger ranges of the heat values (see Figure 5-B) until the OC is able to successfully recognize a best scoring object in these ROIs or until the entire heat range has been exhausted. This flexibility to retry classifications with different sized ROIs gives STAM-Heat a distinct advantage over STAM-Tmin and STAM-Basic approaches to dynamically adapt to the environment, as will be seen in the results section.

Algorithm 2 Discriminate Clustered OOI

```
Given a new STAM M:
For all gradient slicing threshold ranges T:
  Generate ROI_T based on T masking M
    Generate OC scores for ROI_T
Return OOI type from max OC score
```

The OC in this work is based on Google's Inception model which is trained on ImageNet. Specialized training for typical laboratory objects was added to the final layer of the Inception model by using TensorFlow along with a set of training images of local laboratory objects. This approach allowed the object classifier to have a more accurate and relevant classification capability in the application environment than would otherwise generally be supported by the ImageNet dataset training alone.

C. Head Pose Tracking

The attention modeling technique in this work fundamentally requires an accurate head pose tracking system in order for the OA to accurately render the 3D STAMs. The two types of trackers used in this work were selected based on the type of experiments explored. These include an OOI team search task in a 3D environment and a clustered OOI discrimination task in a one-on-one tabletop environment.

For the OOI search task, it was desired that the subject head pose tracker work from the fore and aft point of views in addition to operating from a single sensor perspective in order to facilitate a self-contained system for OA mobility. These characteristics were not found in any available affordable tracking systems, thus a custom technique was implemented. A safety helmet marked with

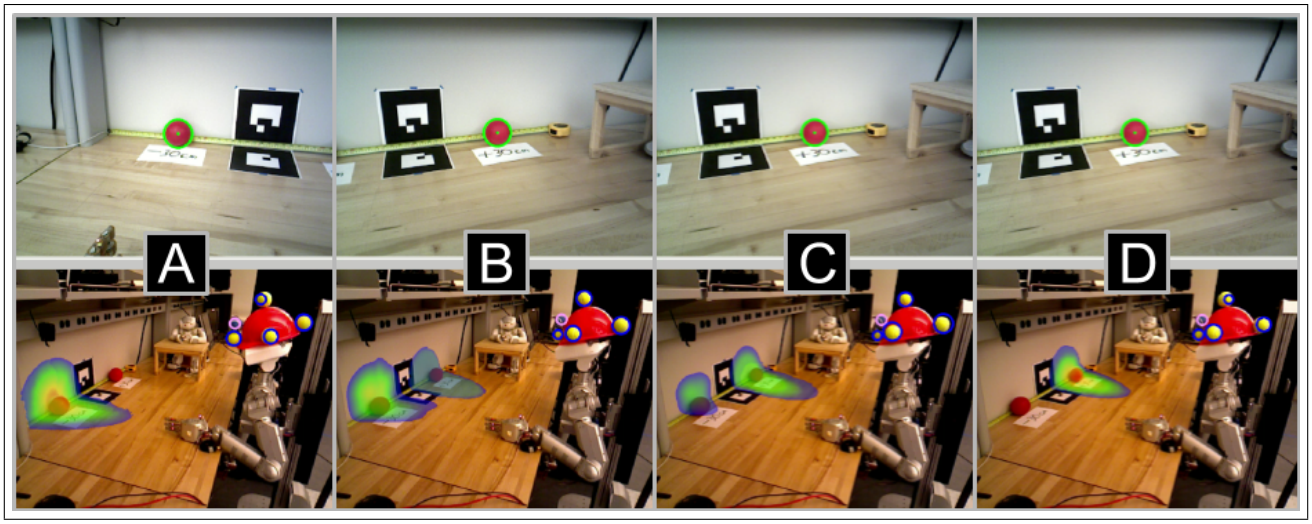

Fig. 3. Model verification apparatus. Top Row: Subject POV. Bottom Row: OA POV.

tennis balls, which serves a similar role to the motion capture bodysuits previously discussed, was worn by the subject and tracked with single RGB-D camera. Locating the marker ball constellation in 3D space with color based image processing techniques via OpenCV allowed the subject's head pose to be localized in 3D space.

For the clustered tabletop OOI discrimination task, an alternative head pose tracking technique can be used because the one-on-one shared workspace scenario does not require OA mobility or aft perspective of the subject. In this task, the user wears a pair of glasses with an attached camera instead of the helmet marker. This allows the subject and OA head poses to be localized based on a table fiducial marker via ARPose. This technique shows that our implementation of STAMs are agnostic about the pose tracker type and also has the advantage of leveraging the fiducial tracking software, foregoing the helmet marker apparatus, and providing real time verification of the subject's Point Of View (POV). However, as an anchoring limitation, it requires that both the OA and subject simultaneously keep the table marker within view of the cameras at all times. Both of these tracking systems are low cost of around $200 US.

D. Validation of the Attention Models

Figure 3 shows the bench test apparatus used to validate the correct operation of our STAM implementation and quantify the accuracy of the helmet tracking system. A Meka M1 humanoid robot served as a measurable human subject replacement in order to provide ground truth measurements for the head pose positions. Robot Operating System (ROS) was used to simultaneously record the OA and M1 sensor data into a rosbag file for analysis. The top row of images shows the M1's POV and bottom row shows the OA's POV.

The humanoid M1 subject, donned with the tracking helmet, was scheduled to periodically visually target OOIs located in a measured table top grid system. The OA generated the 3D model of the subject's attention within the test environment and the model was analyzed for accuracy. It was validated that our STAM implementation accurately reflects the subject's OOIs. For example, Figure 4 shows that during a 30° sweep of the head yaw axis, the helmet tracker maintains an average tracking error of 1.62° as compared to ground truth of the M1's axis sensor. Similar results were observed for the pitch axis. Also note that the STAM accurately lands on the subject's targeted red ball OOI and fades over time in Figure 3-A to Figure 3-D as the OOI target changes from left to right. The accuracy for the glasses camera relies on the ARPose fiducial marker implementation.

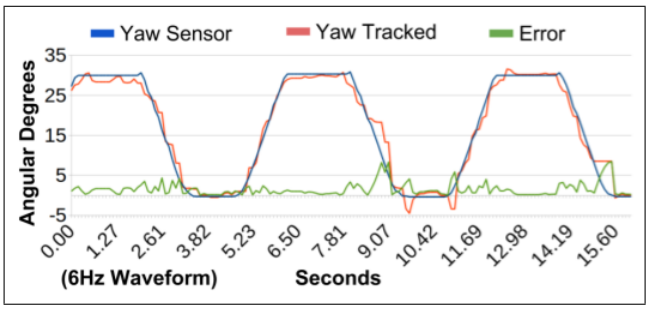

Fig. 4. Model verification apparatus head pose yaw tracking error.

IV. EXPERIMENTS

STAMs are particularly well suited for human-robot team search missions due to their ability to determine a peer's OOIs and unsearched space. Two types of experiments are conducted in order to evaluate the merits of three types of STAMs and better quantify these capabilities. These include a clustered OOI discrimination task in a one-on-one tabletop environment and an OOI search task in a 3D environment.

A. Clustered Objects of Interest discrimination

The first experiment challenges the system to discriminate among clustered OOIs in a one-on-one shared workspace. The human subject periodically focuses on a set of items on the table while the system works to determine the current OOIs, locations, and attention state. Six objects are placed

at 60° increments around a fiducial marker and randomly ordered using Python. This experiment is meant as a pre-requisite to a task such as selecting items for grasp in a human-robot team assembly task. For example, if the OA notices that tool items are of interest, it may pre-calculate grasp path solutions to the items in anticipation of a request for help from the subject. Additionally, it may anticipate the next task or stage of assembly based on the tools or items that are of interest.

The tabletop conditions along with the fiducial marker allow for ARPose to be used for head pose tracking via a set of user worn camera glasses. Additionally, the user's camera POV is available for OOI verification in real time. Gradient slicing of STAM-Heat is used in order to provide ROIs to the OOI classifier. This allows for dynamic scoping of the attention field in order to discriminate clustered OOIs.

B. Object of Interest Search Task in 3D Space

The second experiment is a human-robot team OOI search task that involves a single human search team participant and one robot OA teammate. As shown from the OA's perspective in Figure 1, the OOIs in this experiment are a set of green bottles placed randomly in a simple 10x10m experimentation room. The participant is asked to don the helmet marker and then enter the room to explore for OOIs. The targets are found and scanned by the participant by taking a short video with a camera phone, ostensibly for metering and logging the OOIs. The experiment concludes once the subject determines that all the OOIs have been discovered and scanned or once 5 minutes have passed.

During the experiment, the OA executes the STAM software that observes the human participant. Once the OA can determine the OOI, it contributes to the search mission by highlighting the OOIs, marking their location and labeling the discovery state, as shown in Figure 6. OOIs currently being viewed by the subject are marked in red, ones undiscovered by the subject are marked in blue, and previously discovered OOIs by the subject are marked in green. The OA tracks the subject's OOIs, determines the historical list of OOIs along with their locations, duration of interest, and notes other similar OOIs in the scene. For performance comparison, each of the STAM methods are also applied to this task.

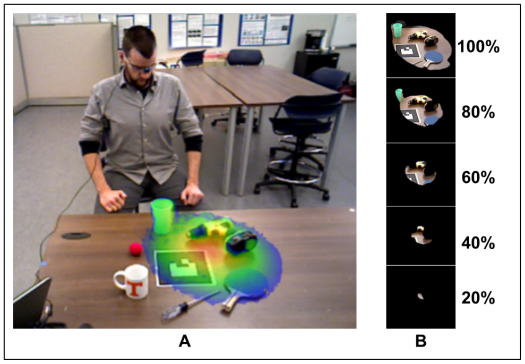

Fig. 5. A: OA POV for clustered OOI task. B: ROI GS Scopes.

V. RESULTS

The results for these experiments show how three STAM techniques and GS perform in two different situations for facilitating implicit communication in a human-robot team.

A. Clustered Objects of Interest Discrimination

Figure 5 shows an example from the clustered OOI discrimination task along with the extracted ROIs from the STAM-Heat model. The GS STAM results for a single experiment of the clustered OOI discrimination task are located in Table I. The columns show the different ROI threshold levels and the row groups show the resultant classification scores for when the user focuses on the different objects. Per Algorithm 2, the top classification scores for each current OOI are marked by the black stars for STAM-Heat. The STAM-Basic and STAM-Tmin methods, marked by the black diamonds, use the entire gaze envelope and are thus in the 100% thresholding level column.

TABLE I
CLASSIFICATION SCORES FOR ONE CLUSTERED OOI EXPERIMENT.

Current OOI	OOI classification scores for ROI gradient slicing percentages during clustered OOI experiment				
Objects	**20%**	**40%**	**60%**	**80%**	**100%**
green cup	0.4655	0.7983	0.9730 ★	0.8727	0.8569 ♦
UTK cup	0.1306	0.1548	0.0253	0.0912	0.1025
paddle	0.1318	0.0100	0.0003	0.0072	0.0100
screw driver	0.1271	0.0011	0.0000	0.0004	0.0004
red ball	0.1033	0.0192	0.0002	0.0010	0.0013
drill	0.0418	0.0166	0.0012	0.0275	0.0289
green cup	0.0038	0.0012	0.0186	0.0080	0.0080
UTK cup	0.8060	0.9341 ★	0.1682	0.3195	0.3195
paddle	0.0074	0.0007	0.0059	0.0071	0.0071
screw driver	0.0213	0.0008	0.0012	0.0016	0.0016
red ball	0.0133	0.0003	0.0002	0.0019	0.0019
drill	0.1482	0.0630	0.8059	0.6619	0.6619 ♦
green cup	0.0005	0.0044	0.7651	0.9860 ★	0.9360 ♦
UTK cup	0.0011	0.0466	0.1623	0.0057	0.0187
paddle	0.9661	0.9264	0.0322	0.0004	0.0055
screw driver	0.0042	0.0073	0.0063	0.0009	0.0026
red ball	0.0277	0.0091	0.0056	0.0001	0.0003
drill	0.0004	0.0062	0.0285	0.0069	0.0369
green cup	0.0018	0.0169	0.0305	0.1866	0.3322
UTK cup	0.0226	0.7187 ★	0.3815	0.0320	0.0334
paddle	0.0210	0.1576	0.0667	0.2015	0.0186
screw driver	0.4039	0.0089	0.1868	0.0097	0.0118
red ball	0.4596	0.0076	0.0118	0.0029	0.0008
drill	0.0912	0.0903	0.3227	0.5673	0.6033 ♦
green cup	0.0001	0.0014	0.2991	0.0587	0.0438
UTK cup	0.0005	0.2070	0.3547	0.2143	0.1097
paddle	0.0010	0.0634	0.1500	0.1155	0.0525
screw driver	0.0001	0.0060	0.0140	0.0108	0.0261
red ball	0.9982 ★	0.7076	0.0886	0.5031	0.6394 ♦
drill	0.0001	0.0146	0.0935	0.0976	0.1286
green cup	0.0004	0.0019	0.0220	0.0120	0.0059
UTK cup	0.0033	0.0463	0.0131	0.0139	0.0161
paddle	0.0014	0.0020	0.0049	0.0031	0.0049
screw driver	0.0919	0.1511	0.0982	0.0218	0.0099
red ball	0.0071	0.0180	0.0033	0.0006	0.0017
drill	0.8958	0.7807	0.8585	0.9486	0.9615 ★♦

♦ - *Indicates choice for STAM-Basic & STAM-Tmin*
★ - *Indicates choice for STAM-Heat*

The clustered OOI discrimination task requires heavy use of the Gradient Slicing technique in order to correctly parse the subject's OOI by providing the object classifier with an ROI tightly bounded around the OOI. STAM-Basic and STAM-Tmin are unable to use dynamic attention scoping once the model is built and thus suffer from dependency on the parameter for the subject's gaze envelope size. If the gaze envelope is too large, these methods tend to cover multiple objects within the space and make it unclear for the classifier to determine which clustered object is actually the correct OOI. The Table II confusion matrix for 20 different clustered OOI experiments highlights this issue because it shows larger objects like the drill and green cup tend to be predicted for this case.

In contrast, STAM-Heat paired with GS performs quite well in this task. For example, the Table III confusion matrix it is able to predict the subject's drill, red ball and paddle OOIs with 100%, 95%, and 85% respectively. The accuracy for the cup OOIs fell to 75%, in part due to the image classifier's propensity to confuse the two cup objects types. The detection of the screwdriver OOI performed poorly overall and seems to indicate a limitation of the classifier to discern the opaque acrylic handle and slender shaft in this environment.

TABLE II

CONFUSION MATRIX FOR 20 CLUSTERED OOI EXPERIMENTS USING STAM-TMIN/STAM-BASIC.

Predicted	Actual					
	green cup	UTK cup	paddle	screw driver	red ball	drill
green cup	85%	20%	55%	20%	20%	0%
UTK cup	5%	45%	10%	25%	50%	0%
paddle	0%	5%	15%	20%	0%	0%
screw driver	0%	0%	0%	5%	0%	0%
red ball	0%	0%	0%	0%	5%	0%
drill	10%	30%	20%	30%	25%	100%

TABLE III

CONFUSION MATRIX FOR 20 CLUSTERED OOI EXPERIMENTS USING STAM-HEAT WITH GRADIENT SLICING.

Predicted	Actual					
	green cup	UTK cup	paddle	screw driver	red ball	drill
green cup	75%	0%	5%	25%	0%	0%
UTK cup	15%	75%	5%	15%	5%	0%
paddle	5%	5%	85%	30%	0%	0%
screw driver	0%	0%	0%	5%	0%	0%
red ball	5%	20%	0%	5%	95%	0%
drill	0%	0%	5%	20%	0%	100%

Table IV shows the GS threshold selection matrix for the different objects over the 20 clustered OOI experiments. This table shows a correlation between the threshold level and the OOI size, which indicates that the GS technique is dynamically adjusting the ROI scope to match the subject's current OOI. For example, the larger drill and green cup OOIs tend to use a large threshold around 60% or greater whereas the small red ball object uses a low threshold of 20%. As the results show, the dynamic ROI scoping technique aids the accuracy of the image classifier by reducing non-salient information in the scene.

TABLE IV

GRADIENT SLICING THRESHOLD SELECTION MATRIX FOR 20 CLUSTERED OOI EXPERIMENTS.

	Threshold				
OOI Type	**20%**	**40%**	**60%**	**80%**	**100%**
green cup	0%	10%	55%	10%	25%
UTK cup	25%	60%	10%	5%	0%
paddle	70%	25%	5%	0%	0%
screw driver	75%	10%	0%	15%	0%
red ball	100%	0%	0%	0%	0%
drill	0%	10%	55%	15%	20%

B. Object of Interest Search Task in 3D Space

Figure 6 shows examples from the human-robot team OOI search task with different STAM modes in each column. The bottom row shows the extracted ROIs from the above STAM images. The left image column shows a gray attention map from STAM-Tmin such that it simply reflects the current gaze envelope of the subject. The middle column shows a gray STAM-Basic built by the subject's gaze envelope over time; however, it does not represent thermal attention scoring areas within the map and thus cannot determine newer from older areas of attention within the map. The right column image shows STAM-Heat with GS.

In Figure 6, the subject enters the scene from the right hand side and moves to sample the two bottles on the table. The bottle in the chair is occluded from the user by the table top and remains undiscovered during the experiment. The OA labels undiscovered OOIs in blue, currently observed OOIs in red and previously discovered OOIs in green. As the STAM-Basic and STAM-Heat versions show, the attention map grows and fades over time in order to represent the change in attention scoring. The STAM-Heat in the right image shows the first bottle cooling and the second bottle heating up as the subject moves to sample a new OOI.

The results for this experiment, as rendered by the OOI location and discovery state labels, indicate that each of the STAM methods is able to determine the OOI type, assist with searching for other items in the environment, and determine the discovery state of these OOIs. This is due in part to the sparse locations of the bottles. These results demonstrate a scenario where the accumulated attention scoring and gradient slicing methods are not needed, as contrasted with the results from the clustered OOIs discrimination task.

Figure 7 shows an OOI discrimination experiment using the bench calibration apparatus where the subject periodically focuses on a red ball and green bottle OOI. This shows an example where the system is able to correctly discriminate the OOIs from an aft perspective of the subject.

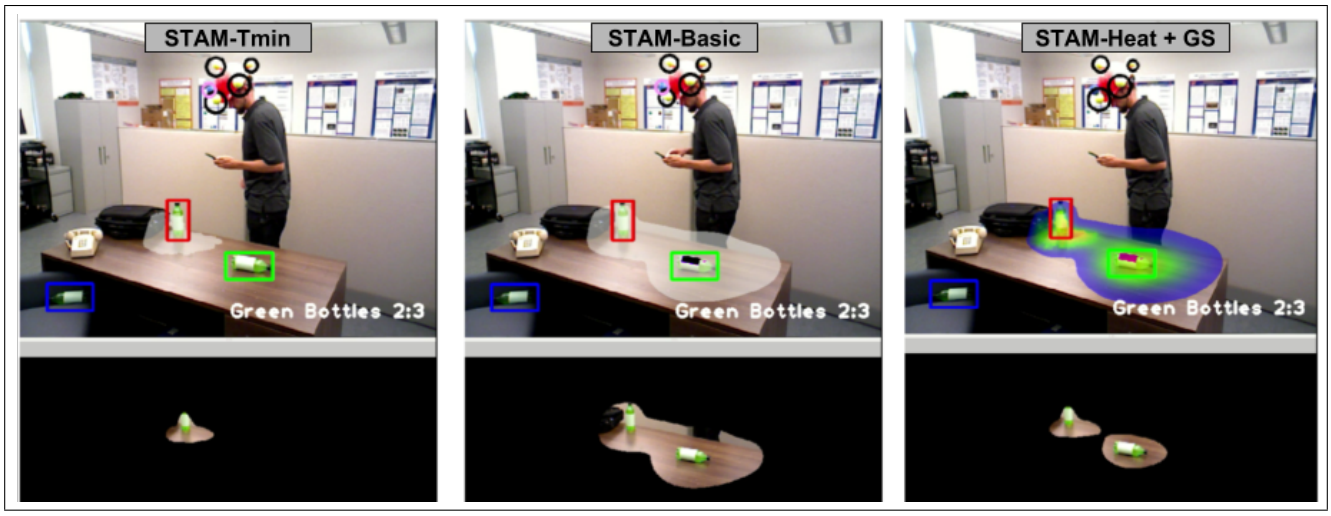

Fig. 6. Top Row: STAMs. Bottom Row: Extracted ROIs.

VI. DISCUSSION

This work is our initial exploration of the 3D STAM concept for use in human-robot teams. Future investigations of this spatio-temporal attention modeling technique could include multi-subject models, multi-OA model integration, highly motile environments, non-human subjects, dynamic hyper-parameters, eye pose deviation, and human trials.

Numerous challenges to implementing this system for real-world environments and applications were encountered in this work. Simple use of a raw 3D point cloud leads to rendering fault conditions due to unscanned objects in the environment. The fix for this would be to use the OA sensor data to build a virtual environment that automatically completes surfaces undetected by the OA sensor. This would require significant prior world knowledge about how to complete real-world scenes and is thus non-trivial. Additionally, no tracking systems or techniques that work well from the aft perspective on unmarked human subjects for acquiring accurate head pose are currently known. Until this issue is solved, STAM implementations for mobile robot teams will likely require human teammates to be marked for tracking.

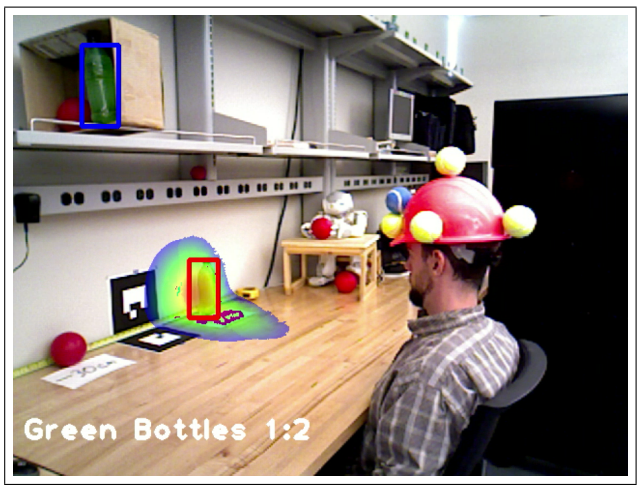

Fig. 7. Aft perspective in OOI discrimination task.

VII. CONCLUSION

This work explores three spatio-temporal attention modeling techniques based on head pose trajectory tracking. We show that STAMs can be used to model implicit information about a human subject's attention and determine OOIs in certain human-robot team applications. Additionally, STAM-Heat paired with the Gradient Slicing technique for OOI discrimination in clustered environments is determined to have classification advantages over STAM-Basic and STAM-Tmin. This work discusses some of the challenges encountered during implementation of STAMs for real-world applications and poses possible solutions to these issues. Finally, multiple ideas for future investigations are proposed.

REFERENCES

[1] P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: A review," Computer Vision and Image Understanding, vol. 141, pp. 152–165, 2015.
[2] Zhang, H. "3D Robotic Sensing of People: Human Perception, Representation and Activity Recognition", PhD Dis., U. of TN, Aug. 2014.
[3] K. Sakita, K. Ogawam, S. Murakami, K. Kawamura, and K. Ikeuchi, "Flexible cooperation between human and robot by interpreting human intention from gaze information," IROS, 2014.
[4] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 185–207, 2013.
[5] A. T. Duchowski, *Eye tracking methodology: theory and practice.* Cham: Springer, 2017.
[6] E. Potapova, M. Zillich, and M. Vincze, "Survey of recent advances in 3D visual attention for robotics," The International Journal of Robotics Research, vol. 36, no. 11, pp. 1159–1176, 2017.
[7] "CUshop," Package InSight. [Online]. Available: https://www.packageinsight.com/cushop/. [Accessed: 30-Apr-2018].
[8] I. Fasel, G. Deak, J. Triesch, and J. Movellan, "Combining embodied models and empirical research for understanding the development of shared attention," ICDL 2002.
[9] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Which one? Grounding the referent based on efficient human-robot interaction," RO-MAN, 2010.
[10] E. Murphy-Chutorian and M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 607–626, 2009.